IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 16, 2021, accepted November 3, 2021, date of publication November 8, 2021, date of current version November 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126338

Progressive Guided Fusion Network With Multi-Modal and Multi-Scale Attention for RGB-D Salient Object Detection

JIAJIA WU[1,2], GUANGLIANG HAN[1], HAINING WANG[3], HANG YANG[1], QINGQING LI[1,2], DONGXU LIU[1,2], FANGJIAN YE[4], AND PEIXUN LIU[1]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Police Administration, People's Public Security University of China, Beijing 100038, China
[4]Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China

Corresponding authors: Guangliang Han (hangl@ciomp.ac.cn) and Peixun Liu (liupx@ciomp.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61602432 and Grant 61401425.

ABSTRACT The depth map contains abundant spatial structure cues, which makes it extensively introduced into saliency detection tasks for improving the detection accuracy. Nevertheless, the acquired depth map is often with uneven quality, due to the interference of depth sensors and external environments, posing a challenge when trying to minimize the disturbances from low-quality depth maps during the fusion process. In this article, to mitigate such issues and highlight the salient objects, we propose a progressive guided fusion network (PGFNet) with multi-modal and multi-scale attention for RGB-D salient object detection. Particularly, we first present a multi-modal and multi-scale attention fusion model (MMAFM) to fully mine and utilize the complementarity of features at different scales and modalities for achieving optimal fusion. Then, to strengthen the semantic expressiveness of the shallow-layer features, we design a multi-modal feature refinement mechanism (MFRM), which exploits the high-level fusion feature to guide the enhancement of the shallow-layer original RGB and depth features before they are fused. Moreover, a residual prediction module (RPM) is applied to further suppress background elements. Our entire network adopts a top-down strategy to progressively excavate and integrate valuable information. Compared with the state-of-the-art methods, experimental results demonstrate the effectiveness of our proposed method both qualitatively and quantitatively on eight challenging benchmark datasets.

INDEX TERMS RGB-D, salient object detection, multi-modal and multi-scale attention, progressive guided fusion.

I. INTRODUCTION

Salient object detection (SOD) aims to locate and segment the interesting or attractive regions in a scene by imitating the human visual system. It has been applied to various vision tasks, such as image segmentation [1], matching [2], enhancement [3], and weakly supervised learning [4]. With the development of depth sensors, depth cues can be conveniently acquired as a supplement to color appearance information, which contributes to better perceive and understand the complex and challenge scenes, such as ones with similar-looking objects and backgrounds, or varying illuminations. Therefore, RGB-D salient object detection using depth cues has attracted more and more attention from researchers.

For a given set of RGB-D (RGB + depth) images, the purpose of RGB-D SOD is to predict a saliency map and extract salient regions by exploring the complementary relationship between color image and depth data. Traditional RGB-D saliency detection models based on hand-crafted features mainly use depth information to excavate a few effective auxiliary feature attributes, such as longitudinal distance, boundaries, shape, surface normal, etc. These properties can improve the ability of models to detect salient objects from complex scenes. Over the past few years, numerous traditional RGB-D models have been developed [5]–[24]. Specifically, some methods focus on taking depth feature as an explicit supplementary of color feature [5]–[16]. For

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

example, Cheng *et al.* [8] extend the 2D center bias to 3D spatial bias via using longitudinal depth distance, and combine it with color contrast and depth contrast for calculating a saliency map. Zhu and Li [13], [14] directly employ the depth map to generate the depth feature saliency and merge it with the color feature saliency, then background elimination model or the center dark channel prior is applied to optimize the fusion saliency map. Others devote themselves to design the depth measurement algorithm to obtain implicit attributes such as shape and contour in the depth map [17]–[21]. In [17], instead of using absolute depth distance, Ju *et al.* propose an anisotropic center-surround difference (ACSD) measure which pops out salient objects from the scenes with the help of global depth structure. In [18], Ren *et al.* present the normalized depth prior and the global-context surface orientation prior. These prior can highlight near objects, weaken distant objects, and decrease the saliency of severely inclined surfaces (such as the ground plane or ceilings). Given that the background contains high-contrast areas may be cause false positives for detection, Feng *et al.* [19] design a local background enclosure (LBE) feature to capture the spread of angular directions, which quantifies the proportion of the object boundary that is in front of the background from the depth map. In general, the depth feature-based method implements RGB-D saliency detection in an intuitive and simple way, while ignores the potential feature attributes in depth map. By contrast, the depth measurement-based method aims to refine the saliency result by utilizing implicit attributes. Moreover, to deal with the varying depth quality, Cong *et al.* [22] present a depth confidence measure to assess the reliability of the depth map and control the fusion ratio of depth and color features.

However, due to the limited expression of traditional handcrafted features, the complementarity between RGB and depth features cannot be fully explored, which greatly astricts the improvement of algorithm performance. To solve the problem, many researchers gradually introduce convolutional neural network (CNN) for better integrating RGB and depth data [25]–[40]. The CNN-based models can learn deeper feature representations, profoundly excavate the associations between RGB images and depth cues to improve the saliency detection performance. According to the fusion strategy, RGB-D saliency detection networks can be roughly divided into three types [41]: 1) early fusion; 2) late fusion; 3) multi-scale fusion. The early fusion directly connects RGB image and depth map to form a four-channels as input [32], or first combines low-level features extracted from the independent networks and then feeds it to the subsequent network [25]. The late fusion mainly adopts two parallel networks to learn high-level features from RGB and depth maps, respectively, and then connect them to generate the final saliency prediction map [40]. The multi-scale fusion mainly integrates cross-modal interactive module into the feature learning networks and explores the complementarity between deep-layer and shallow-layer features, which is also the most mainstream fusion strategy at present. Typically, Hao *et al.* [26] propose a multi-scale multi-path fusion network, which diversifies the single fusion path into a global reasoning one and a local capturing one, and meanwhile introduces multi-level cross-modal interactions in multiple layers to achieve sufficient and efficient fusion. Li *et al.* [37] design an information conversion module to integrate high-level RGB and depth features in an interactive and adaptive way, and a cross-modal depth-weighted combination block to enhance RGB features with depth features at each level. Moreover, Wang *et al.* [42] propose a completely different fusion method from the above-mentioned networks. To prevent the dual-stream architecture from preferring RGB sub-branch in the subsequent fusion process, they design a novel data-level recombination strategy which converts the original four-dimensional RGB-D data into DGB, RDB and RGD. Then, these reorganized data are sent to a lightweight three-stream network for complementary fusion.

It has been proved that the depth map with rich spatial information is meaningful to detect salient objects from a cluttered background. However, due to the limitations of depth sensor, the quality of depth maps will vary greatly in different environments. The poor depth map often tremendously endures serious noise or blurred edges, which severely affects the detection accuracy, and even leads to detection failure. In response to this problem, some interesting works [43], [44] have emerged. Wang *et al.* [43] propose a simple yet effective D (depth) quality measurement scheme. The core idea is to design a series of features based on the common attributes of high-quality D regions, then, combine them with RGB and D saliency cues to guide selective RGB-D fusion. Chen *et al.* [44] present a two-stage depth estimation method. First, the corresponding relationship between the input and its similar images is established through retrieval, and it combines with the designed depth transferring strategy to estimate the coarse depth. Then, they construct fine-grained, object-level correspondences coupled for improving the quality of estimated depth, and finally feed the estimated depth and original depth into the selective fusion network. However, these depth measurements and estimations will cause additional calculations and time.

On the whole, although traditional and deep learning-based approaches have achieved good results, how to further alleviate the impact of poor depth maps and effectively integrate RGB and depth modalities is still a challenge worth exploring. Therefore, based on the above observations, we further clarify the main purpose of this RGB-D SOD task, which is to design an effective and universal fusion network. The network is capable of excelling in extracting salient objects without additional depth measurement schemes, regardless of the depth quality in the scene. In other words, according to the provided RGB images and depth images with unknown quality, the network can adaptively complete the valuable and complementary fusion action rather than biasing towards a certain modality attribute or a fixed and stiff ratio fusion. To achieve this goal, in this work, we present a progressive guided fusion network (PGFNet) with multi-modal and

multi-scale attention for RGB-D salient object detection. The PGFNet is composed of four key parts. First, we adopt two parallel ResNet-50 [45] or VGG-16 [46] to extract RGB and depth features, respectively. Next, a multi-modal and multi-scale attention fusion model (MMAFM) is designed for adaptively fusing multiple modal features in each layer. Then, we propose a multi-modal feature refinement mechanism (MFRM) to optimize the original RGB and depth features with the assistance of high-level fusion feature. At last, the residual prediction module (RPM) is used to predict the saliency map of each layer. We alternately cascade these modules in top-down manner, which continuously enhances and optimizes the fusion of multi-modal features to obtain the final saliency prediction map.

The main contributions of our paper can be summarized as follows:

1) We structure a novel network, i.e., PGFNet, which aims to adequately and efficiently learn the complementarity of multi-modal features and multi-scale features in diverse layers, as well as detect salient regions more accurate.

2) We design a multi-modal and multi-scale attention fusion model (MMAFM), which utilizes the semantic associations between modalities to adaptively fuse features at different modalities and different scales for selecting and enhancing valuable information.

3) To better express the semantic information of multiple modalities, we propose a multi-modal feature refinement mechanism (MFRM), which combines the high-level fusion feature to further optimize the shallow-layer features, so that they can retain more details while having richer global context information.

4) Comprehensive experiments on eight popular benchmark datasets under five widely used evaluation metrics demonstrate that the proposed PGFNet is pretty competitive to the state-of-the-art RGB-D salient object detection models.

The rest of this article is organized as follows: In Section II, we elaborate the proposed PGFNet. In Section III, we conduct extensive experiments to confirm the superiority and effectiveness of our PGFNet. In Section IV, we provide the conclusion.

II. METHODOLOGY

In this section, we elaborate the proposed progressive guided fusion network (PGFNet) with multi-modal and multi-scale attention. As illustrated in Fig. 1, the overall framework is based on a symmetrical two-stream encoder-decoder architecture, which is mainly consists of four subsections: feature encoding module, multi-modal and multi-scale attention feature fusion model, high-level fusion feature guided multi-modal feature refinement mechanism and residual prediction module.

A. FEATURE ENCODING

Considering the computational complexity, we employ ResNet-50 [45] network for feature encoding. As shown in Fig. 1, RGB image and depth map are encoded separately through the two-stream encoders. To be concise, we denote the encoding block in the RGB branch as $E_R^i (i \in \{1, 2, 3, 4, 5\}$, i is the block index), the corresponding output feature as F_R^i, define the encoding block in the depth branch as $E_D^i (i \in \{1, 2, 3, 4, 5\})$, the corresponding out feature as F_D^i, Given an RGB image I_R and an aligned depth map I_D, through the encoding blocks, we obtain two feature groups $\mathbf{F}_R = \{F_R^1, F_R^2, F_R^3, F_R^4, F_R^5\}$ and $\mathbf{F}_D = \{F_D^1, F_D^2, F_D^3, F_D^4, F_D^5\}$, each of which contains five features with different levels and diverse scales. The values in each encoding block represent the length, width, and channel size of the output feature, respectively. When we adopt ResNet-50 as the backbone, the input resolution of RGB image and depth map are set to $352 \times 352 \times 3$ and $352 \times 352 \times 1$, respectively.

B. MULTI-MODAL AND MULTI-SCALE ATTENTIVE FUSION MODEL

For an RGB-D SOD task, how to effectively utilize depth cues is a crucial point. An accurate depth map can provide precise spatial structure clues and promote the detection accuracy. In contrast, a poor depth map contains massive disturbance and error information, which is detrimental to the detection performance. Therefore, how to adequately aggregate RGB and depth cues at different layers is extremely critical. Inspired by stereoscopically attentive multi-scale (SAM) module [47], we propose a multi-modal and multi-scale attentive fusion model (MMAFM), as shown in Fig. 2. Different from the SAM module, our model not only adaptively learns the weight factor of each scale according to the characteristic of own modality, but also globally guides the selection and optimization at the corresponding modal scale by combining the multi-modal information. In detail, MMAFM is composed of three processes: multi-scale feature extraction, multi-modal and multi-scale attention, feature fusion, which are explained in detail below.

1) MULTI-SCALE FEATURE EXTRACTION

In order to acquire more plentiful global semantic information and reduce the information dilution in the decoder, a multi-scale operation is applied to deep-layer features. Specifically, as illustrated on the left part of Fig. 2, the portion contains five parallel branches at each modality. For all branches, a 1×1 convolution is adopted to compress the channel size to 32, which greatly reduce the calculation and complexity of the network model. Then, four parallel 3×3 dilated convolutions with different dilation rate are applied to obtain abundant global context. Without loss of generality, for the i^{th} layer refined features $\tilde{F}_R^i$ and $\tilde{F}_D^i$ $(i \in \{1, 2, 3, 4, 5\})$, after the multi-scale operation, we get the multi-scale feature groups of RGB and depth, i.e., $\mathbf{f}_R = [f_R^0, f_R^1, \ldots, f_R^k]$ and $\mathbf{f}_D = [f_D^0, f_D^1, \ldots, f_D^k]$, which is described as,

$$f_{R/D}^k = \begin{cases} Conv\left(\tilde{F}_{R/D}^i\right), & k = 0 \\ DConv\left(Conv\left(\tilde{F}_{R/D}^i\right)\right), & k = 1, 2, \ldots, N-1 \end{cases} \quad (1)$$

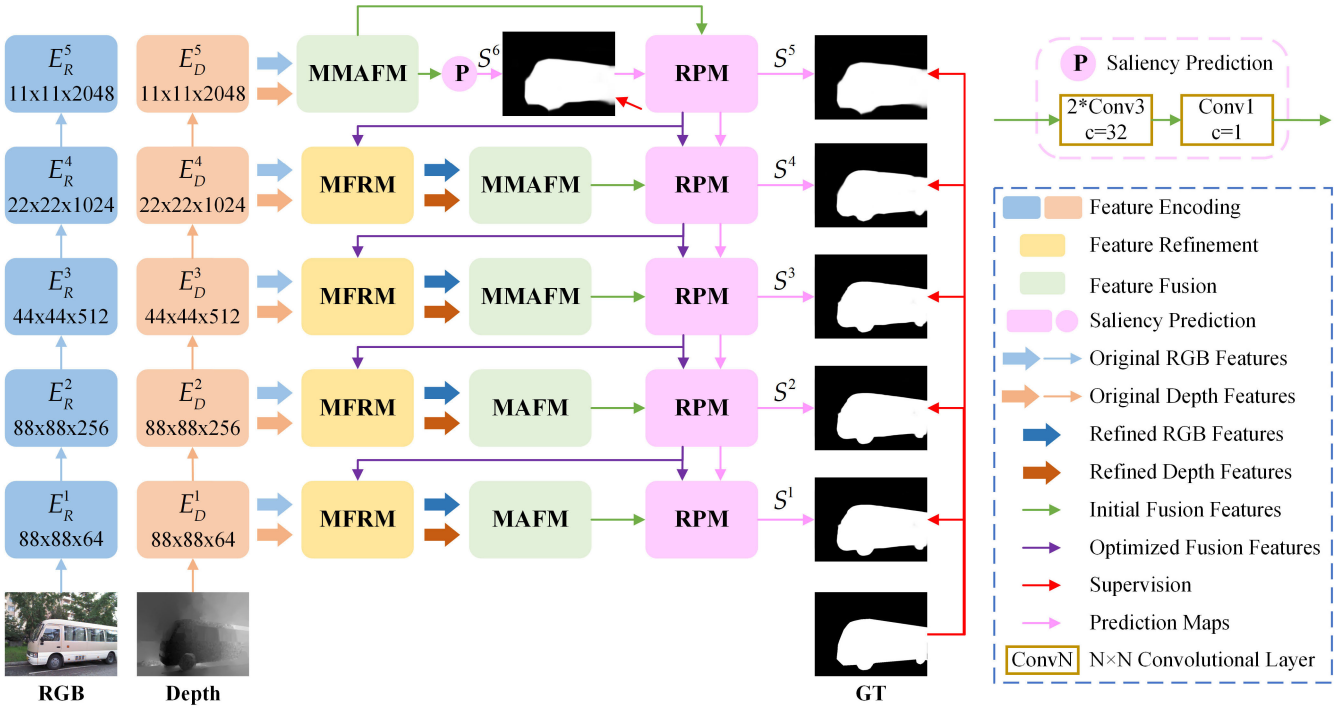


FIGURE 1. The overall architecture of our proposed PGFNet, which consists of four key stages: feature encoding, feature fusion, feature refinement and saliency prediction. First, the feature encoding networks (ResNet-50 [45]) extract features from RGB and depth images. Next, the multi-modal features of the highest layer are fed to multi-modal and multi-scale attention fusion model (MMAFM) for adaptive integration and enhancing the response to beneficial features. Then, the features of the remaining layers will be transmitted to a high-level guided multi-modal feature refinement mechanism (MFRM) before fusion, pursuing more details and global context information. Finally, the residual prediction modules optimize and decode the fusion features in each layer for highlighting salient objects. Here, MAFM refers a multi-modal attention fusion model without multi-scale information, which is applied at the shallow layers. Notably, at the training phase, the pixel-level ground truth (GT) is used to supervise all saliency maps generated by the network.

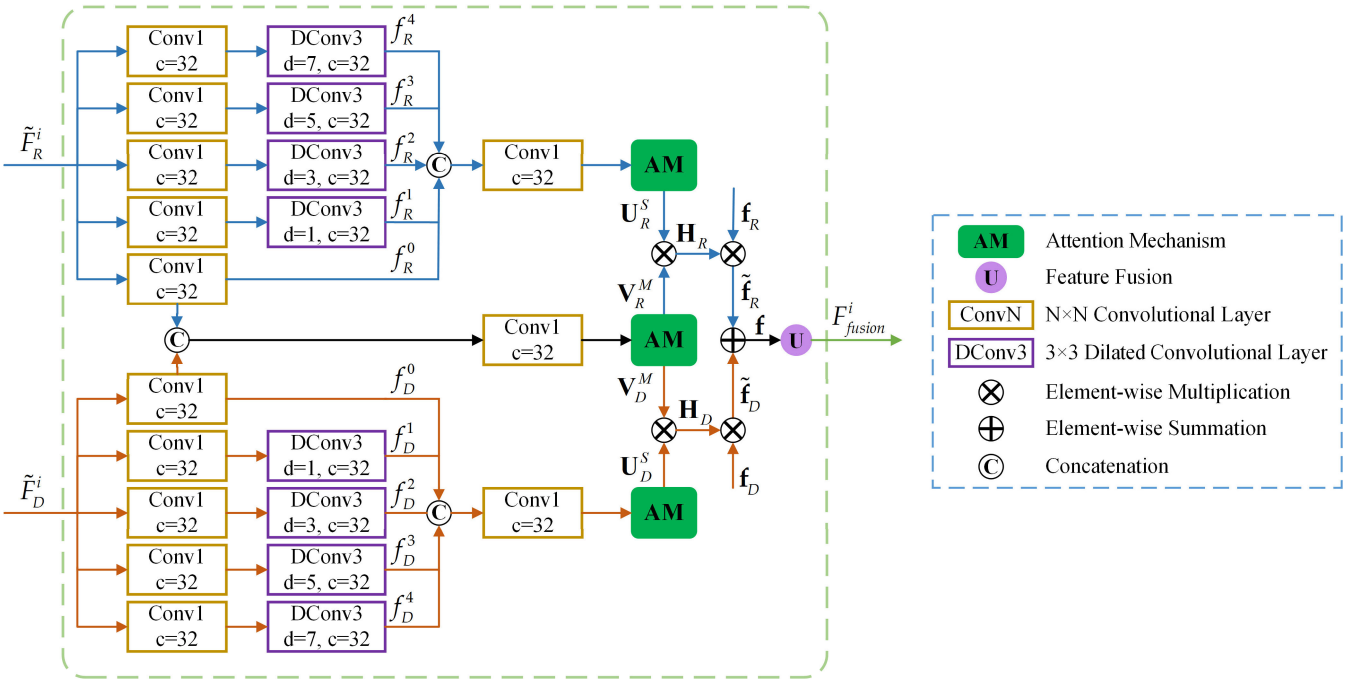


FIGURE 2. The overall structure of MMAFM. Parameters c, d represent the output channel size and dilation rate, respectively.

where, $\tilde{F}_R^i$ and $\tilde{F}_D^i$ are the refined RGB and depth features which are described in detail in Section II-C. k represents the branch index, N is the total number of branches. The more branches, the greater computation is required. So, N is

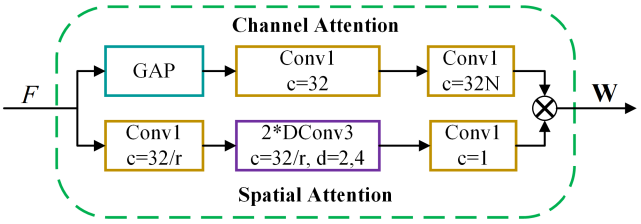


FIGURE 3. Illustration of attention mechanism (AM). The parameter r represents the reduction factor, which is set to 4.

empirically set to 5 in this article. $Conv(\cdot)$ represents a 1×1 convolution operation, and $DConv(\cdot)$ represents a 3×3 dilated convolution, for the k^{th} branch ($k \in \{1, 2, 3, N-1\}$), its dilation rate is $2k-1$. Considering the computational complexity and the sallow-layer features (i.e., $\tilde{F}_R^i$ and $\tilde{F}_D^i$, $i \in \{1, 2\}$) may be contain more noise, the multi-scale operation is not applied to these layers, that is, k here is only equal to 0.

2) MULTI-MODAL AND MULTI-SCALE ATTENTION

If all the scale features of above RGB and depth modalities are directly aggregated by simple element-wise summation or connection, it may cause the beneficial information branches to be weakened or even submerged by the useless ones. In addition, the information provided by each branch may have a different focus, which is often overlooked if they are treated equally. To alleviate this issue, the modal attention and scale attention are combined into the fusion model. We not only consider the importance of each scale feature in own modality, but also integrate and utilize the information of multiple modalities to adaptively guide and refine the scale feature at each modality. By this way, adequately explore the complementarity and difference of features from the aspect of modalities and scales. In detail, as shown on the right part of Fig. 2, we separately connect all branches of RGB and depth modalities, and feed them into the attention mechanisms (AM) after a 1×1 convolution operation for obtaining the scale attention weights $\mathbf{U}_R^S = [u_R^0, u_R^1, \ldots, u_R^k]$ and $\mathbf{U}_D^S = [u_D^0, u_D^1, \ldots, u_D^k]$ at the corresponding modality. Subsequently, connect the outputs of all 0^{th} branch and input it into the attention mechanisms after a 1×1 convolution for getting the modal attention weights $\mathbf{V}_R^M = [v_R^0, v_R^1, \ldots, v_R^k]$ and $\mathbf{V}_D^M = [v_D^0, v_D^1, \ldots, v_D^k]$ at the corresponding scale. As shown in Fig. 3, the attention mechanism includes channel attention and spatial attention. It aims to learn the attention of each branch by using all branches of a single modality or the original features of all modalities, suppress non-informative features and focus on the specific spatial location in a global manner. The operation of the attention mechanism $\mathcal{F}_{AM}(\cdot)$ is generally defined as,

$$\mathbf{W} = \mathcal{F}_{AM}(F) = \mathcal{F}_{CA}(F) \otimes \mathcal{F}_{SA}(F) \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{N \times C \times H \times W}$ includes the attention weights of N branches, $\mathbf{W} = \left[w^0, w^1, \ldots, w^k\right] (k \in \{0, 1, \ldots, N-1\})$, w^k is the attention weight of the k^{th} branch. $F \in \mathbb{R}^{C \times H \times W}$ is the input feature, H, W, C represent its length, width and channel size. $\mathcal{F}_{CA}(\cdot)$ and $\mathcal{F}_{SA}(\cdot)$ represent the channel attention and spatial attention, respectively. $\otimes$ denotes the element-wise multiplication. More specifically,

$$\mathcal{F}_{CA}(F) = \mathcal{F}_{MLP}(\mathcal{F}_{GAP}(F)) \tag{3}$$

where $\mathcal{F}_{GAP}(\cdot)$ represents the global average pooling (GAP) operation, $\mathcal{F}_{MLP}(\cdot)$ is a multi-layer (two-layer) perceptron. And the spatial attention is defined as,

$$\mathcal{F}_{SA}(F) = Convs(F) \tag{4}$$

where $Convs(\cdot)$ represents the execution of four convolution operations in sequence: a 1×1 convolution, two 3×3 dilated convolutions and a 1×1 convolution. Next, we can formally calculate the scale attention weights $\mathbf{U}_R^S$, $\mathbf{U}_D^S$ and the modal attention weights $\mathbf{V}_R^M$, $\mathbf{V}_D^M$ according to (2),

$$\mathbf{U}_{R/D}^S = \mathcal{F}_{AM}\left(Conv\left(Cat\left(f_{R/D}^0, f_{R/D}^1, \ldots, f_{R/D}^k\right)\right)\right) \tag{5}$$

where $\mathbf{U}_R^S = [u_R^0, u_R^1, \ldots, u_R^k]$ includes the RGB scale weights extracted from all RGB branches, u_R^k is the scale attention weight of the k^{th} branch at RGB modality. Similarly, $\mathbf{U}_D^S = [u_D^0, u_D^1, \ldots, u_D^k]$ contains the depth scale weights extracted from all depth branches, u_D^k is the scale attention weight of the k^{th} branch at depth modality. $Cat(\cdot)$ means connection operation, $Conv(\cdot)$ represents a 1×1 convolution operation for reducing parameters and computational complexity.

$$\mathbf{V}_{R/D}^M = \mathcal{F}_{AM}\left(Conv\left(Cat\left(f_R^0, f_D^0\right)\right)\right) \tag{6}$$

where $\mathbf{V}_R^M = [v_R^0, v_R^1, \ldots, v_R^k]$ includes the RGB modal weights under different RGB scales, which is learned from all original modal features (i.e., RGB and depth modalities). Similarly, $\mathbf{V}_D^M = [v_D^0, v_D^1, \ldots, v_D^k]$ contains the depth modal weights under different depth scales. Then, we calculate the final weight of each branch combined with (5) and (6).

$$\mathbf{H}_{R/D} = \mathbf{U}_{R/D}^S \otimes \mathbf{V}_{R/D}^M \tag{7}$$

where $\mathbf{H}_R = [h_R^0, h_R^0, \ldots, h_R^k]$, $\mathbf{H}_D = [h_D^0, h_D^0, \ldots, h_D^k]$ are the compositive attention weights of RGB and depth branches, respectively. $h_{R/D}^k = u_{R/D}^k \otimes v_{R/D}^k$ denotes the attention weight of the k^{th} RGB or depth branch. Then, the weighted multi-scale features $\tilde{\mathbf{f}}_R$ and $\tilde{\mathbf{f}}_D$ can be described as,

$$\tilde{\mathbf{f}}_{R/D} = \mathbf{H}_{R/D} \otimes \mathbf{f}_{R/D} \tag{8}$$

where $\tilde{\mathbf{f}}_R = [\tilde{f}_R^1, \tilde{f}_R^2, \ldots, \tilde{f}_R^k]$, $\tilde{\mathbf{f}}_D = [\tilde{f}_D^1, \tilde{f}_D^2, \ldots, \tilde{f}_D^k]$. $\tilde{f}_{R/D}^k$ means the weighted scale feature of the k^{th} RGB or depth branch, $\tilde{f}_{R/D}^k = h_{R/D}^k \otimes f_{R/D}^k$.

3) FEATURE FUSION

For the multi-scale feature fusion at multiple modalities, we adopt a two-step way: inter-modal fusion and inter-scale fusion. First, we simply aggregate the modal scales from the same branches by element-wise summation.

$$\mathbf{f} = \tilde{\mathbf{f}}_R \oplus \tilde{\mathbf{f}}_D \tag{9}$$

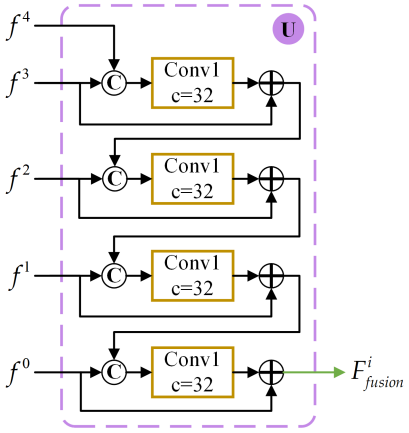


FIGURE 4. Detailed diagram of inter-scale feature fusion.

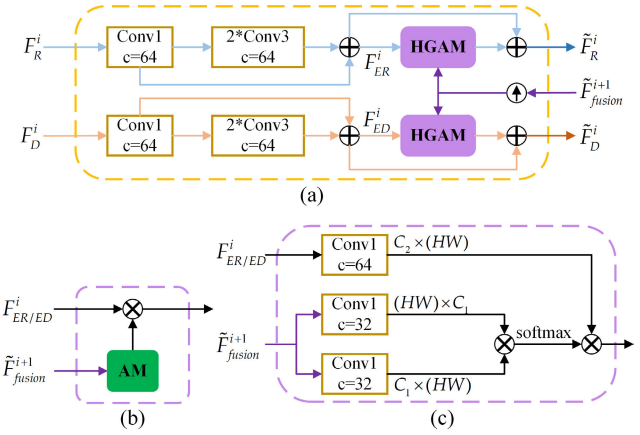


FIGURE 5. Illustration of MFRM. (a) The overall structure of MMRM. (b) High-level guided attention mechanism without self-attention. (c) High-level guided attention mechanism with self-attention.

where $\mathbf{f} = [f^0, f^1, \ldots, f^k]$ is the multi-scale feature group after executing inter-modal fusion, $f^k = \tilde{f}^k_R \oplus \tilde{f}^k_D$ is the feature merged by all the k^{th} modality branch, and $\oplus$ denotes the element-wise summation. It is worth noting that dilated convolutions can enlarge the receptive field of features, whereas large dilation rate may lead to serious gridding effect and lose more spatial details. To overcome it, we conduct a top-down progressive fusion on the scale levels, as shown in Fig. 4. Normally, the smaller dilation rate, the less information continuity is destroyed, and the more original details can be retained. Therefore, we try to continuously propagate the high-level scale feature with large dilation rate to the low level, so that the final fusion feature can be supplemented with contextual information while reserving local details. Concretely, the fusion operation between two adjacent scale levels can be expressed as,

$$f^{k,k+1} = f^k + Conv\left(Cat\left(f^k, f^{k+1,k+2}\right)\right) \tag{10}$$

where $k \in \{0, 1, \ldots, N-2\}$, $f^{k,k+1}$ represents the fusion feature between the k^{th} and $(k+1)^{th}$ scale levels. When $N = 5$, the initial fusion feature $f^{N-1,N} = f^{N-1} = f^4$. Then, we can calculate the final fusion feature at the i^{th} layer by (10), i.e., $F^i_{fusion} = f^{0,1}$.

C. MULTI-MODAL FEATURE REFINEMENT MECHANISM

The output features from deep-layer encoding blocks contain rich high-level semantic information, which can indicate the approximate position and shape of the object in the scene. To this end, we design a multi-modal feature refinement mechanism (MFRM) guided by the high-level feature. Its main goal is to gradually supplement the deep-layer multi-modal fusion feature with strong semantics to the shallow-layer original features by the top-down cascade. The proposed refinement strategy can effectively improve the global semantic representation ability of each modal feature, reduce the interference of redundant information, automatically select and strengthen important feature cues for saliency detection, as well as further promote the quality of the subsequent fusion operation. As shown in Fig. 5a, we adopt a symmetrical attention sub-module to capture the relationship between each individual modality and the high-level fusion feature. Specifically, for the RGB feature F^i_R of the i^{th} ($i \in \{1, 2, 3, 4\}$) layer, we first feed it into a 1×1 convolution to reduce the channel size to 64. Two 3×3 convolution blocks are followed to enlarge the receptive field and extract more available unimodal information. In addition, to prevent the information loss during the convolution process, the compressed feature is added by a residual connection for retaining more original attribute. The same operation is also applied to F^i_D for strengthening the depth feature. Then, we obtain the enhanced RGB and depth features (i.e., F^i_{ER} and F^i_{ED}), and input them together with the upper-layer optimized fusion feature $\tilde{F}^{i+1}_{fusion}$ into the high-level guided attention mechanism (HGAM) for separately learning supplementary enhancement features. It is worth noting that $\tilde{F}^{i+1}_{fusion}$ is the fusion feature optimized by the residual prediction module (RPM), and the specific calculation process, see Section II-D. The main structure of HGAM is shown in Fig. 5b. Furthermore, inspired by the successful application of self-attention [48]–[50], we refer to it in the refined model to replace the model in Fig. 5b, as shown in Fig. 5c. Considering the computational complexity of self-attention, it is only applied to the 3^{th} and 4^{th} layers. The operations of the two attention mechanisms are marked as $\mathcal{F}_{HGAM1}$ and $\mathcal{F}_{HGAM2}$, respectively, then we have,

$$\tilde{F}^i_{R/D} = \begin{cases} F^i_{ER/ED} + \mathcal{F}_{HGAM1}\left(F^i_{ER/ED}, \tilde{F}^{i+1}_{fusion}\right), & i \in \{1, 2\} \\ F^i_{ER/ED} + \mathcal{F}_{HGAM2}\left(F^i_{ER/ED}, \tilde{F}^{i+1}_{fusion}\right), & i \in \{3, 4\} \end{cases} \tag{11}$$

where, $\tilde{F}^i_R$ and $\tilde{F}^i_D$ represent the refined RGB and depth features, respectively, which are employed as the input features of the multi-modal and multi-scale attention fusion model (MMAFM). Notably, $\tilde{F}^5_{R/D} = F^5_{R/D}$, which means that the top layer is fused directly without a refinement mechanism.

Mathematically,

$$\mathcal{F}_{HGAM1}\left(F^{i}_{ER/ED}, \tilde{F}^{i+1}_{fusion}\right) = F^{i}_{ER/ED} \otimes \mathcal{F}_{AM}\left(U\left(\tilde{F}^{i+1}_{fusion}\right)\right) \tag{12}$$

where $U(\cdot)$ denotes the up-sampling operation via bilinear interpolation if these features are not in the same scale. The process of the $\mathcal{F}_{HGAM2}$ can be described as follows,

$$F_{in} = \mathcal{F}_{UP}\left(\tilde{F}^{i+1}_{fusion}\right) \tag{13}$$

$$w_A = softmax\left(\left(R_1\left(Conv(F_{in})\right)\right)^{\top} \otimes R_1\left(Conv(F_{in})\right)\right) \tag{14}$$

$$\mathcal{F}_{HGAM2}\left(F^{i}_{ER/ED}, \tilde{F}^{i+1}_{fusion}\right) = R_2\left(R'_1\left(Conv\left(F^{i}_{ER/ED}\right)\right) \otimes w_A^{\top}\right) \tag{15}$$

where F_{in} is the feature after up-sampling, and w_A is a attention weight which considers the pairwise relationship at any point in the high-level feature map. *softmax* $(\cdot)$ is an element-wise softmax function, $R_1(\cdot)$ reshapes the input feature to $\mathbb{R}^{C_1\times(HW)}$, $R'_1(\cdot)$ reshapes the input feature to $\mathbb{R}^{C_2\times(HW)}$, and $R_2(\cdot)$ reshapes the input feature to $\mathbb{R}^{C_2\times H\times W}$. Notably, C_1 is set to 1/2 of $C_2 = C$.

D. RESIDUAL PREDICTION MODULE

As we all know, the shallow layers of deep learning networks capture low-level structural cues while the deep layers capture high-level semantic information. To take maximum advantage of the complementarity and difference between diverse feature layers, our network construction has been adopting a progressive guided manner to integrate and refine features. It is committed to gradually transfer high-level semantic information from the deep layer to the lower layer and learn a more accurate salient object with clearer edges. Furthermore, we can obtain a rough saliency map through the top-layer multi-modal fusion feature, the map can indicate the approximate location and shape of salient object, meanwhile effectively suppressing and eliminating most of background elements. Based on the above observations, we design a residual prediction module (RPM) based on the saliency map to further optimize the shallow-layer fusion features by combining with the deep-layer saliency cues. This operation can tellingly alleviate the gradual sparseness of high-level information during the fusion process as well as suppress the background noise from the low-level features. As illustrated in Fig. 6, input an initial fusion feature F^{i}_{fusion} of the i^{th} layer, an optimized fusion feature $\tilde{F}^{i+1}_{fusion}$ and a prediction map S^{i+1} of the $(i+1)^{th}$ layer, it outputs the optimized fusion feature $\tilde{F}^{i}_{fusion}$ and prediction map S^{i},which are defined as:

$$S^{i+1}_{up} = \delta\left(U\left(S^{i+1}\right)\right) \tag{16}$$

$$\tilde{F}^{i}_{fusion} = F^{i}_{fusion} + U\left(\tilde{F}^{i+1}_{fusion}\right) + Conv(S^{i+1}_{up} \times F^{i}_{fusion}) \tag{17}$$

$$S^{i} = S^{i+1}_{up} + Conv\left(\tilde{F}^{i}_{fusion}\right) \tag{18}$$

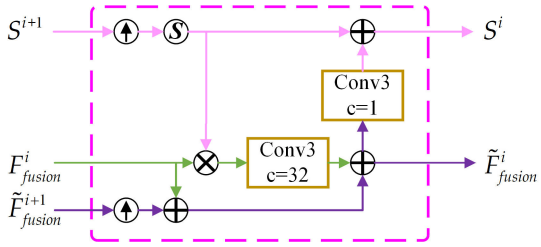


FIGURE 6. Detailed diagram of the residual prediction module (RPM).

where S^{i+1}_{up} is the saliency prediction map after up-sampling, $\delta(\cdot)$ is a sigmoid function. The outputs of this layer will guide the multi-modal feature refinement module and residual prediction module of the next layer, and so on, until the first layer. For the top layer, we first utilize the initial fusion feature F^{5}_{fusion} to directly generate an initial prediction map, i.e., S^{6}, and then feed them to the residual prediction module.

E. LOSS FUNCTION

As shown in Fig. 1, we supervise the prediction map at each layer, which clarifies the optimization goal for each step of the network and accelerates the convergence of training. Moreover, to better guide the network learning and produce more details, we introduce a pixel position aware (PPA) loss [51], which synthesizes local structure information to generate different weights for all pixels and introduces pixel constraints (i.e., weight binary cross entropy (wBCE) loss) and global constraints (i.e., weighted intersection over union (wIoU) loss). Mathematically,

$$L_{PPA} = L_{wBCE} + L_{wIoU} \tag{19}$$

where L_{wBCE} is defined as,

$$L_{wBCE} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(1+\gamma\alpha_{ij})\sum_{l=0}^{1}\mathbf{1}(GT_{ij}=l)\log\mathbf{Pr}(S_{ij}=l|\Psi)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\gamma\alpha_{ij}} \tag{20}$$

where, where H and W are the height and the width of the saliency map. $\mathbf{1}(\cdot)$ is the indicator, and γ is a hyperparameter. $l \in \{0, 1\}$ denotes two classes of the labels. S_{ij} and GT_{ij} are prediction saliency map and ground truth of the pixel at location (i, j) in an image. Ψ represents all the parameters of the model, and $\mathbf{Pr}(S_{ij} = l|\Psi)$ is the predicted probability. α_{ij} is the indicator of pixel importance, which is defined by the difference between the center pixel and its surroundings. L_{wBCE} is beneficial for the model to pay more attention to hard edge areas. Homoplastically, the wIoU loss is defined as,

$$L_{wIoU} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(GT_{ij} * S_{ij}) * (1+\gamma\alpha_{ij})}{\sum_{i=1}^{H}\sum_{j=1}^{W}(GT_{ij} + S_{ij} - GT_{ij} * S_{ij}) * (1+\gamma\alpha_{ij})} \tag{21}$$

In summary, the total loss function of our network is expressed as,

$$L = \sum_{i=1}^{6} \alpha_i L_{PPA}\left(S^i, GT\right) \tag{22}$$

where α_i is the weight coefficient and simply set to 1 in our experiments.

III. EXPERIMENTS

A. DATASETS

To demonstrate the effectiveness of our proposed method, we evaluate it on eight public benchmark datasets.

NJU2K [17] contains 1985 RGB-D images which are collected from the Internet, 3-D movies and photographs taken by stereo camera, and depth maps are estimated by the optical-flow method.

NLPR [9] includes 1000 RGB-D images, where the depth maps are captured by Microsoft Kinect.

STERE [5] contains 1000 pairs of binocular images with the corresponding pixel-level ground truth. This is the first collection of stereoscopic images in this field.

DES [8] is a small dataset comprises 135 indoor RGB-D images, taken by Kinect at a resolution of 640 × 640.

SSD [13] is built on three stereo movies and includes indoor and outdoor scenes. It has 80 images with the resolution of 960 × 1080.

SIP [30] consists of 929 high-resolution images, which designed for salient person detection in the complex scenes. The depth maps are captured by a smart phone (Huawei Mate10).

LSDF [52] includes 100 light fields captured by a Lytro light field camera.

DUT [29] is a recently released dataset containing 800 indoor and 400 outdoor scenes, some of which are quite challenging.

B. EVALUATION METRICS

Following [30], we use the following five popular evaluation metrics to comprehensively evaluate the performance of the saliency detection methods.

MAE estimates a mean absolute error between a prediction saliency map S and a ground-truth map GT, it is defined as

$$\mathrm{MAE} = \frac{1}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x, y) - GT(x, y)| \tag{23}$$

PR curve is formed by a series of pairs of precision and recall scores calculated at fixed thresholds ranging from 0 to 255, which describes the model performance at different situations.

$$precision = \frac{|S \cap GT|}{S} \tag{24}$$

$$recall = \frac{|S \cap GT|}{GT} \tag{25}$$

F-measure is a harmonic mean of average precision and recall, which is defined as,

$$F_\beta = \frac{(1 + \beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{26}$$

we empirically set $\beta^2 = 0.3$.

S-measure [53] is used to measure the spatial structure information, which is defined as,

$$S_\alpha = \alpha \cdot S_0 + (1 - \alpha) \cdot S_r \tag{27}$$

where α is a balance parameter between the object-aware structural similarity S_0 and region-aware structural similarity S_r, and it is set to 0.5. S_r is the sum of the structural similarity of multiple image blocks with different weights. The greater the proportion of blocks covering GT foreground region, the greater the weight allocated. S_0 is the comprehensive consideration of foreground structure similarity and background structure similarity.

E-measure [54] is to evaluate the foreground map (FM) and noise, which combines local pixel values with image-level mean values to jointly capture image-level statistics and local pixel matching information.

$$E_m = \frac{1}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_{FM}(x, y) \tag{28}$$

where ϕ is an enhanced alignment matrix for the two properties of a binary map.

C. EXPERIMENTAL SETTINGS

1) TRAINGING/TESTING

Following the same training settings as most models, such as in [25], [30], [55], we employ 1485 images from the NJU2K dataset and 700 images from the NLPR dataset as our training set. The remaining images in the NJU2K and NLPR datasets and the whole datasets of STERE, DES, SSD, SIP, and LSDF are used for testing.

2) IMPLEMENTATION DETAILS

We adopt the PyTorch [56] framework to build our network model and conduct extensive experiments on an NVIDIA TITAN Xp GPU. The feature encoder is composed of two parallel ResNet-50 networks. The networks discard the last pooling and fully connected layers, the parameters are initialized according to the pre-training model of ImageNet [57]. The other parameters in our network are initialized as the PyTorch default settings. Refer to [58], we utilize the Adam algorithm [59] to optimize our model. The initial learning rate is set to 0.0001, and it drops to 0.1 times every 60 epochs with a total of 200 epochs. The images are resized to 352 × 352 for both the training and testing stages. For augmenting the training samples, we also take some measures (i.e., random flipping, rotating, and clipping). It takes about 12 hours to train our model with batch size of 8.

TABLE 1. Quantitative results of ablation studies on three popular datasets. Red indicates the best performance, ↑ denotes larger is better, and ↓ denotes smaller is better. BCE: binary cross entropy loss. PPA: pixel position aware loss. MMAFM: multi-modal and multi-scale attention fusion model, where −1/−0 represents using or not using top-down multi-scale feature fusion strategy, respectively. MFRM: multi-modal feature refinement mechanism, where −1/−0 indicates with or without the self-attention, respectively. RPM: residual prediction module.

	Components							NJU2K [17]				NLPR [9]				SIP [30]			
	BCE	PPA	MMAFM-0	MMAFM-1	MFRM-0	RPM	MFRM-1	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$
1	√							0.8850	0.9343	0.8761	0.0568	0.8922	0.9496	0.8622	0.0383	0.8483	0.9071	0.8385	0.0741
2		√						0.8856	0.9353	0.8789	0.0513	0.8955	0.9513	0.8672	0.0328	0.8492	0.9087	0.8411	0.0685
3		√	√					0.9087	0.9489	0.9090	0.0375	0.9167	0.9574	0.9038	0.0260	0.8761	0.9240	0.8827	0.0527
4		√		√				0.9113	**0.9493**	0.9142	0.0356	0.9200	0.9577	0.9100	0.0241	0.8761	0.9217	0.8833	0.0509
5		√		√	√			0.9120	0.9477	0.9131	0.0352	0.9234	0.9624	0.9131	0.0233	0.8785	0.9223	0.8853	0.0508
6		√		√	√	√		0.9125	0.9486	0.9140	0.0351	0.9247	0.9637	0.9140	0.0232	0.8797	0.9243	0.8890	0.0493
7		√		√		√	√	**0.9153**	0.9466	**0.9159**	**0.0341**	**0.9272**	**0.9647**	**0.9185**	**0.0218**	**0.8823**	**0.9262**	**0.8896**	**0.0480**

D. ABLATION STUDY

Our network combines multi-modal and multi-scale attention fusion model (MMAFM), multi-modal feature refinement mechanism (MFRM) and residual prediction module (RPM). In this subsection, we provide comprehensive ablution experiments on NJU2K [17], NLPR [9], and SIP [30] to demonstrate the effectiveness of these components. Table 1 intensively shows all the results of the above experiments. Specifically, we analysis 1) the importance of multi-modal and multi-scale attention fusion model (MMAFM); 2) the necessity of multi-modal feature refinement mechanism (MFRM) and residual prediction module (RPM); 3) the usefulness of PPA loss. We change only one component at a time, leaving the other parameters unchanged. In this paper, we directly connect RGB and depth features of the highest layer extracted from ResNet-50 to predict a saliency map, which is set as the baseline model.

1) THE IMPORTANCE OF MMAFM

The MMAFM plays a very important role in the proposed PGFNet. To study its importance, we explore two variables relative to the baseline model: replacing the top-down fusion strategy in MMAFM with direct element-wise summation (i.e., the 3^{rd} row), MMAFM uses the designed fusion strategy (i.e., the 4^{th} row). As shown in Table 1, compared with the baseline model (i.e., the 2^{nd} row), all evaluation metrics (in the 3^{rd} and 4^{th} rows) obviously show a gradual increase trend. Overall, the proposed MMAFM improves (2.45∼2.69%, 0.64∼1.4%, 3.53∼4.28%, 0.87∼1.76%) for the metrics (S_α, E_m, F_β, M) on three datasets. Conclusively, the results of the 3^{rd} and 4^{th} rows confirm that the top-down fusion strategy is better than the direct summation operation, the reason may be that this way can better alleviate the gridding effect and reserve more spatial details. With the help of MMAFM, our PGFNet captures a more efficient semantic representation of salient objects by taking full advantage of the complementarity between RGB and depth features in terms of different scales and modalities.

2) THE NECESSITY OF MFRM AND RPM

To verify the necessity of MFRM and RPM in our PGFNet, we provide three variables based on the baseline model combine with MMAFM (i.e., the 4^{th} row): introducing MFRM without self-attention (i.e., the 5^{th} row), joining MFRM and RPM without self-attention (i.e., the 6^{th} row), adding MFRM and RPM with self-attention (i.e., the 7^{th} row). In Table 1, the 5^{th} row is overwhelmingly better than the 4^{th} row, confirming that our proposed refinement mechanism is helpful to the optimization of salient objects. It is able to reinforce the semantic expression of shallow-layer features. The performance of the 6^{th} row is improved compared with the 5^{th} row, which indicates that our prediction model further promotes the detection accuracy. Finally, we introduce self-attention into the deep-layer MFRM (i.e., the 3^{rd} and 4^{th} layers) to obtain the final result (i.e., the 7^{th} row), which demonstrates that the high-level guided self-attention mechanism is quite valuable and achieves further the optimization of network performance.

3) THE USEFULNESS OF PPA LOSS

To illustrate the usefulness of loss function, we provide two variables: BCE loss (i.e., the 1^{st} row) and PPA loss (i.e., the 2^{nd} row). As shown in Table 1, compared with BCE loss, PPA loss has certain advantages in all evaluation metrics (S_α, E_m, F_β, M) with the increase of about (0.16%, 0.14%, 0.35%, 0.55%), especially in MAE, the result is more prominent. This may be that PPA loss pays more attention to the perception of the edge and structural information of salient objects.

E. COMPARISONS WITH THE STATE-OF-THE-ART

1) COMPARISON METHODS

We compare our proposed algorithm with 20 state-of-the-art RGB-D SOD models, including 10 traditional hand-crafted-features-based methods: LHM [9], DESM [8], ACSD [17], GP [18], LBE [19], DCMC [22], SE [12], CDCP [14], MDSF [24], CDB [16]; and 10 CNN-based methods: DF [25], CTMF [40], PCF [60], AFNet [55], MMCI [26], TANet [31], CPFP [28], D3Net [30], PGAR [38], DQAS [43]. Table 2

TABLE 2. Quantitative comparisons of different RGB-D SOD methods on seven popular datasets. All the CNN-based models are trained on the NJU2K and NLPR datasets. Red, green and blue indicate the best, second and third performances. ↑ denotes larger is better, and ↓ denotes smaller is better.

	Models	Year	NJU2K [17]				NLPR [9]				STERE [5]				DES [8]				SSD [13]				SIP [30]				LFSD [52]			
			$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$
Hand-crafted-Features-Based	LHM [9]	2014	0.5136	0.4471	0.6318	0.2048	0.6298	0.7662	0.6221	0.1077	0.3778	0.7706	0.6832	0.1719	0.5779	0.6526	0.5108	0.1138	0.5663	0.7167	0.5678	0.1951	0.5109	0.7156	0.5745	0.1837	0.5573	0.7699	0.7119	0.2106
	DESM [8]	2014	0.6648	0.7908	0.7169	0.2835	0.5722	0.8047	0.6404	0.3124	0.6425	0.8108	0.6998	0.2951	0.6224	0.8677	0.7652	0.2986	0.6021	0.7694	0.6796	0.3081	0.6157	0.7699	0.6692	0.2981	0.7220	0.8184	0.7661	0.2480
	ACSD [17]	2014	0.6992	0.8028	0.7114	0.2021	0.6728	0.7803	0.6065	0.1787	0.6919	0.8063	0.6688	0.2000	0.7283	0.8498	0.7561	0.1685	0.6753	0.7849	0.6824	0.2028	0.7316	0.8382	0.7633	0.1721	0.7341	0.8366	0.7673	0.1881
	GP [18]	2015	0.5265	0.7029	0.6470	0.2106	0.6545	0.7234	0.6110	0.1461	0.5876	0.7431	0.6708	0.1822	0.6359	0.6702	0.5968	0.1682	0.6148	0.7820	0.7398	0.1798	0.5876	0.7682	0.6873	0.1729	0.6398	0.8318	0.7870	0.1828
	LBE [19]	2016	0.6952	0.8026	0.7477	0.1528	0.7619	0.8545	0.7450	0.0813	0.6601	0.7869	0.6327	0.2498	0.7026	0.8899	0.7883	0.2082	0.6209	0.7364	0.6187	0.2781	0.7272	0.8526	0.7509	0.2004	0.7356	0.8040	0.7262	0.2084
	DCMC [22]	2016	0.6861	0.7987	0.7151	0.1716	0.7244	0.7926	0.6475	0.1167	0.7306	0.8191	0.7403	0.1476	0.7071	0.7726	0.6661	0.1111	0.7042	0.7855	0.7110	0.1689	0.6828	0.7434	0.6183	0.1859	0.7527	0.8564	0.8172	0.1547
	SE [12]	2016	0.6642	0.8128	0.7479	0.1687	0.7561	0.8474	0.7132	0.0913	0.7082	0.8461	0.7548	0.1427	0.7408	0.8557	0.7412	0.0896	0.6751	0.7999	0.7102	0.1653	0.6281	0.7709	0.6606	0.1644	0.6981	0.8395	0.7906	0.1666
	CDCP [14]	2017	0.6685	0.7407	0.6209	0.1803	0.7270	0.8200	0.6451	0.1121	0.7134	0.7864	0.6644	0.1489	0.7092	0.8110	0.6313	0.1145	0.6028	0.7004	0.5345	0.2139	0.5950	0.7211	0.5052	0.2244	0.7173	0.7860	0.7026	0.1665
	MDSF [24]	2017	0.7477	0.8378	0.7752	0.1572	0.8051	0.8847	0.7931	0.0950	0.7281	0.8091	0.7188	0.1762	0.7412	0.8511	0.7462	0.1224	0.6727	0.7793	0.7030	0.1923	0.7169	0.7982	0.6976	0.1669	0.7004	0.8264	0.7828	0.1897
	CDB [16]	2018	0.6239	0.7421	0.6476	0.2028	0.6286	0.7908	0.6181	0.1142	0.6151	0.8227	0.7173	0.1655	0.6452	0.8297	0.7226	0.1004	0.5617	0.6977	0.5921	0.1959	0.5574	0.7366	0.6204	0.1923	0.5203	0.7737	0.6816	0.2181
CNN-Based	DF [25]	2017	0.7626	0.8639	0.8043	0.1410	0.8018	0.8800	0.7775	0.0847	0.7574	0.8475	0.7573	0.1409	0.7522	0.8703	0.7660	0.0933	0.7465	0.8283	0.7351	0.1423	0.6529	0.7587	0.6566	0.1854	0.7906	0.8651	0.8171	0.1382
	CTMF [40]	2018	0.8493	0.9131	0.8447	0.0847	0.8599	0.9290	0.8255	0.0561	0.8480	0.9123	0.8305	0.0863	0.8631	0.9318	0.8441	0.0554	0.7757	0.8648	0.7285	0.0993	0.7158	0.8291	0.6941	0.1394	0.7956	0.8645	0.7915	0.1194
	PCF [60]	2018	0.8768	0.9245	0.8720	0.0592	0.8736	0.9250	0.8410	0.0437	0.8746	0.9247	0.8603	0.0635	0.8418	0.8929	0.8038	0.0491	0.8414	0.8944	0.8074	0.0618	0.8424	0.9006	0.8377	0.0706	0.7942	0.8350	0.7786	0.1121
	AFNet [55]	2019	0.7725	0.8529	0.7748	0.0998	0.7990	0.8793	0.7712	0.0584	0.8246	0.8869	0.8234	0.0750	0.7696	0.8811	0.7288	0.0680	0.7140	0.8074	0.6868	0.1177	0.7203	0.8193	0.7115	0.1178	0.7381	0.8152	0.7355	0.1335
	MMCI [26]	2019	0.8588	0.9150	0.8526	0.0789	0.8557	0.9130	0.8149	0.0591	0.8728	0.9274	0.8630	0.0676	0.8477	0.9282	0.8224	0.0647	0.8133	0.8823	0.7809	0.0820	0.8329	0.8968	0.8179	0.0862	0.7871	0.8386	0.7716	0.1318
	TANet [31]	2019	0.8785	0.9252	0.8741	0.0605	0.8861	0.9409	0.8632	0.0410	0.8712	0.9231	0.8605	0.0596	0.8582	0.9099	0.8275	0.0460	0.8393	0.8973	0.8097	0.0629	0.8347	0.8950	0.8298	0.0751	0.8014	0.8473	0.7958	0.1108
	CPFP [28]	2019	0.8777	0.9227	0.8767	0.0533	0.8884	0.9317	0.8675	0.0359	0.8792	0.9252	0.8738	0.0513	0.8720	0.9235	0.8458	0.0379	0.8067	0.8516	0.7664	0.0817	0.8501	0.9029	0.8501	0.0636	0.8279	0.8716	0.8255	0.0879
	D^{3}Net [30]	2020	0.9005	0.9385	0.8999	0.0463	0.9118	0.9530	0.8972	0.0297	0.8986	0.9385	0.8913	0.0459	0.8979	0.9455	0.8851	0.0310	0.8570	0.9107	0.8346	0.0584	0.8603	0.9086	0.8612	0.0631	0.8251	0.8621	0.8114	0.0948
	PGAR [38]	2020	0.8909	0.9261	0.8852	0.0479	0.9163	0.9488	0.8961	0.0274	0.8942	0.9288	0.8804	0.0447	0.8863	0.9240	0.8644	0.0322	0.8319	0.8721	0.7985	0.0676	0.8384	0.8863	0.8270	0.0726	0.8163	0.8611	0.7985	0.0909
	DQAS [43]	2021	0.8922	0.9283	0.8908	0.0514	0.8998	0.9380	0.8835	0.0342	0.8968	0.9324	0.8880	0.0481	0.8791	0.9310	0.8632	0.0360	0.8254	0.8600	0.7924	0.0756	—	—	—	—	0.8441	0.8837	0.8388	0.0860
Ours	PGFNet	2021	0.9153	0.9466	0.9159	0.0341	0.9272	0.9647	0.9185	0.0218	0.9019	0.9436	0.9010	0.0396	0.9335	0.9720	0.9284	0.0162	0.8682	0.9149	0.8532	0.0485	0.8823	0.9262	0.8896	0.0480	0.8346	0.8735	0.8405	0.0877

and Fig. 7 show the quantitative comparison results of the proposed method on seven datasets. We also report saliency maps with various scenes, as shown in Fig. 8. For a fair comparison, the saliency maps of all compared methods are directly provided by their authors or generated by running their released codes.

Moreover, we additionally provide a comparison with the latest 8 CNN-based methods: DMRA [29], SSF [61], S^{2}MA [62], HDFNet [63], FRDT [35], DANet [36], CoNet [64], A2dele [65]. These methods have the same training set, and the set introduces 800 images from the DUT dataset besides the subsets of NJU2K and NLPR datasets mentioned above. In turn, we retrain our model on this training set and list all the results in Table 3.

2) QUANTITATIVE COMPARISON

We report the PR curves and F-measure curves on seven datasets in Fig. 7 and list S-measure (S_α), maximum E-measure (E_m), maximum F-measure (F_β), MAE (M) in Table 2 and Table 3. As shown in Fig. 7, our method achieves better PR curves and F-measure curves on all datasets. This indicates that the proposed PGFNet can obtain the higher precision and recall compared with other methods, as well as means that the saliency maps we generate have better consistency.

As listed in Table 2, it is obvious that the performance of CNN-based models is far superior to traditional ones, which yet proves the status and application value of convolutional neural network in the image processing field. Undoubtedly, compared with traditional or deep learning-based models, the proposed algorithm shows powerful competitiveness in terms of all evaluation metrics. Performance gains over the best compared models (D^{3}Net, PGAR and DQAS) are (0.3%∼5.4%, 0.4%∼5.5%, 1%∼6.5%, 0.5%∼2.5%) for the metrics (S_α, E_m, F_β, M) on all datasets except LFSD dataset. We only do not achieve the best on the three values of the LFSD dataset, but the values are still sub-optimal. Moreover, from Table 3, we can clearly find that our evaluation data is also excellent on the new training set.

In conclusion, the effectiveness of our model is relatively ideal from a quantitative point of view.

3) VISUAL COMPARISON

We provide visual comparisons with classical four non-deep learning and six CNN-based models in Fig. 8. We observe that the proposed method is able to handle several challenging and complicated scenes. To more convincing, we compare these methods in following aspects: (1) the ability to handle boundary contacts; (2) the ability to resolve similar appearances; (3) the detection ability for a poor depth map. (4) the ability to process a depth with distractors.

Here combine with examples in Fig. 8 to vividly explain the above five aspects. First, in the 4^{th}, 7^{th} and 8^{th} rows, only PGAR method responses well to boundary contact issues. But it may misdetect when the object has a low contrast in the scene, especially in the 8^{th} row. It fails to make full use of the depth map with clear contours and mistakes the object shadow as a salient area. In contrast, our saliency maps perform better in this situation. Second, as shown in the 6^{th} and 8^{th} rows, the object appearance is relatively close

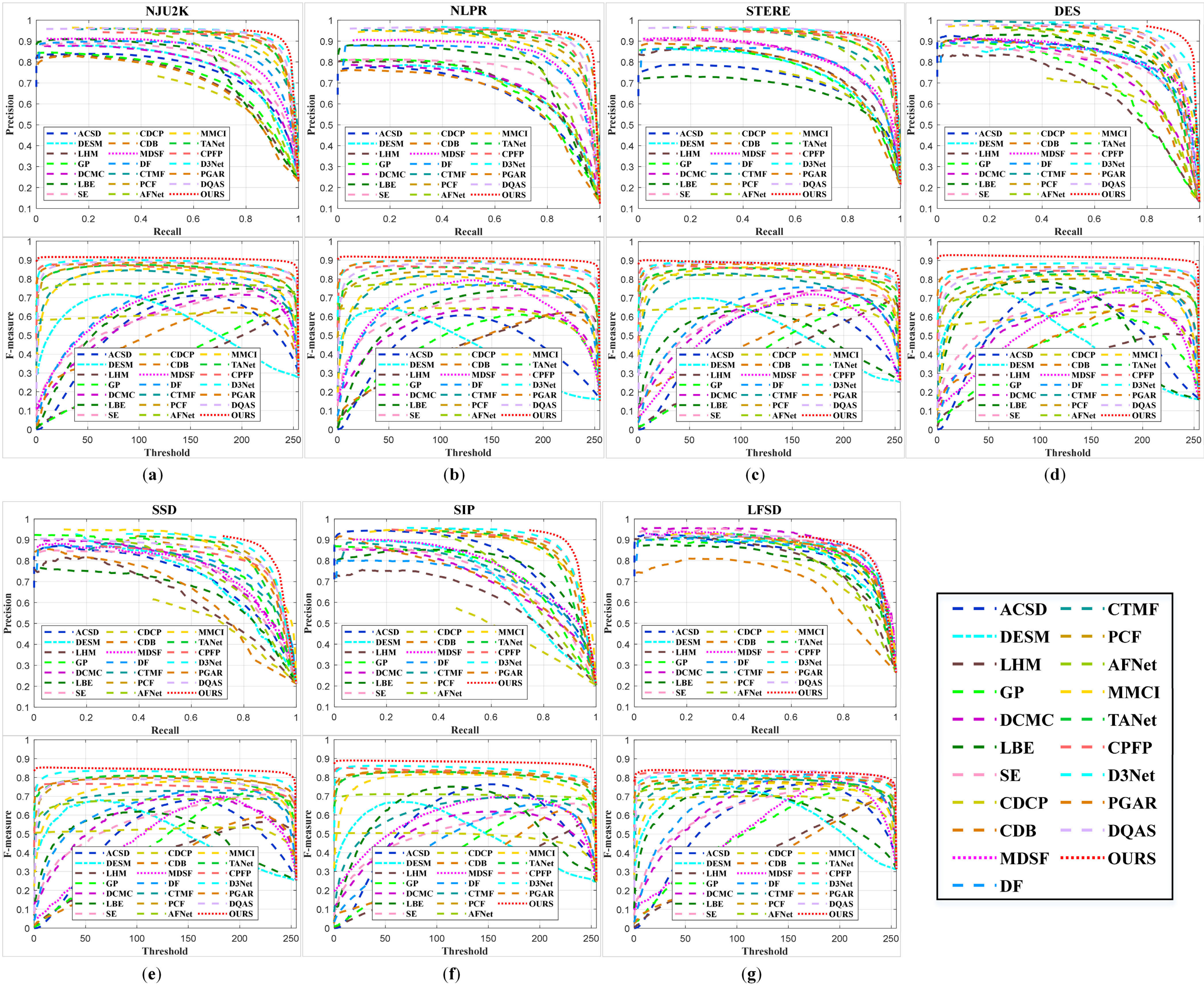


FIGURE 7. Quantitative comparisons of our proposed and 20 state-of-the-art methods on seven datasets. (a) NJU2K dataset. (b) NLPR dataset. (c) STERE dataset. (d) DES dataset. (e) SSD dataset. (f) SIP dataset. (g) LFSD dataset. The first rows are PR curves, and the second rows are F-measure under different thresholds.

TABLE 3. Quantitative comparisons of different RGB-D SOD methods on eight popular datasets. All the models are trained on the NJU2K, NLPR and DUT datasets. Red, green and blue indicate the best, second and third performances. ↑ denotes larger is better, and ↓ denotes smaller is better.

Models	Year	NJU2K [17]				NLPR [9]				STERE [5]				DES [8]			
		$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$
DMRA [29]	2019	0.8863	0.9270	0.8864	0.0506	0.8989	0.9470	0.8795	0.0313	0.8448	0.9157	0.8571	0.0627	0.8996	0.9426	0.8878	0.0300
SSF [61]	2020	0.8564	0.9117	0.8558	0.0636	0.8850	0.9332	0.8616	0.0350	0.8372	0.9120	0.8396	0.0646	0.8631	0.9107	0.8384	0.0381
S^2MA [62]	2020	0.8943	0.9299	0.8888	0.0532	0.9155	0.9532	0.9017	0.0298	0.8904	0.9324	0.8823	0.0507	0.9405	0.9733	0.9347	0.0209
HDFNet [63]	2020	0.9109	0.9446	0.9127	0.0369	0.9159	0.9557	0.9044	0.0268	0.9055	0.9468	0.9075	0.0387	0.9320	0.9710	0.9242	0.0199
FRDT [35]	2020	0.8984	0.9331	0.8986	0.0476	0.9145	0.9500	0.8998	0.0288	0.8926	0.9391	0.8921	0.0459	0.8995	0.9387	0.8862	0.0299
DANet [36]	2020	0.8971	0.9359	0.8927	0.0463	0.9086	0.9487	0.8937	0.0306	0.8921	0.9304	0.8815	0.0475	0.9048	0.9575	0.8947	0.0282
CoNet [64]	2020	0.8955	0.9369	0.8927	0.0461	0.9079	0.9450	0.8871	0.0306	0.9051	0.9471	0.9013	0.0374	0.9093	0.9446	0.8957	0.0283
A2dele [65]	2020	0.8676	0.9138	0.8717	0.0521	0.8896	0.9370	0.8751	0.0309	0.8785	0.9276	0.8794	0.0444	0.8833	0.9196	0.8732	0.0299
Ours	2021	0.9181	0.9521	0.9217	0.0332	0.9235	0.9609	0.9134	0.0236	0.9061	0.9443	0.9029	0.0372	0.9328	0.9712	0.9327	0.0173

Models	Year	SSD [13]				SIP [30]				LFSD [52]				DUT [29]			
		$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$
DMRA [29]	2019	0.8569	0.9063	0.8441	0.0583	0.8059	0.8750	0.8209	0.0854	0.8471	0.9005	0.8565	0.0754	0.8996	0.9426	0.8878	0.0300
SSF [61]	2020	0.7897	0.8676	0.7621	0.0837	0.7989	0.8702	0.7857	0.0913	0.7491	0.8371	0.7609	0.1239	0.8844	0.9261	0.8894	0.0477
S^2MA [62]	2020	0.8684	0.9094	0.8482	0.0522	0.8721	0.9185	0.8771	0.0569	0.8370	0.8728	0.8352	0.0944	0.8847	0.9368	0.9005	0.0435
HDFNet [63]	2020	0.8661	0.9087	0.8486	0.0477	0.8783	0.9224	0.8857	0.0497	0.8471	0.8805	0.8380	0.0851	0.9057	0.9407	0.9086	0.0409
FRDT [35]	2020	0.8716	0.9242	0.8651	0.0533	0.8473	0.8977	0.8523	0.0711	0.8568	0.9032	0.8594	0.0734	0.9103	0.9484	0.9192	0.0385
DANet [36]	2020	0.8692	0.9067	0.8524	0.0499	0.8784	0.9206	0.8840	0.0537	0.8452	0.8860	0.8459	0.0816	0.8893	0.9312	0.8954	0.0472
CoNet [64]	2020	0.8530	0.9145	0.8403	0.0595	0.8581	0.9129	0.8670	0.0628	0.8621	0.9066	0.8595	0.0707	0.9191	0.9563	0.9273	0.0333
A2dele [65]	2020	0.8024	0.8618	0.7764	0.0699	0.8284	0.8896	0.8333	0.0698	0.8328	0.8740	0.8318	0.0768	0.8846	0.9300	0.8921	0.0423
Ours	2021	0.8721	0.9201	0.8584	0.0425	0.8890	0.9316	0.9004	0.0445	0.8582	0.8978	0.8649	0.0703	0.9213	0.9542	0.9262	0.0321

to the background, especially in the 6^{th} row, the color of the chair is quite similar to the door behind it. All CNN-based comparison methods cannot effectively extract the complete object from the scene, while our model achieve the goal by roundly exploring the depth cues, and improve the detection accuracy. Thirdly, the quality of the depth maps is very poor in

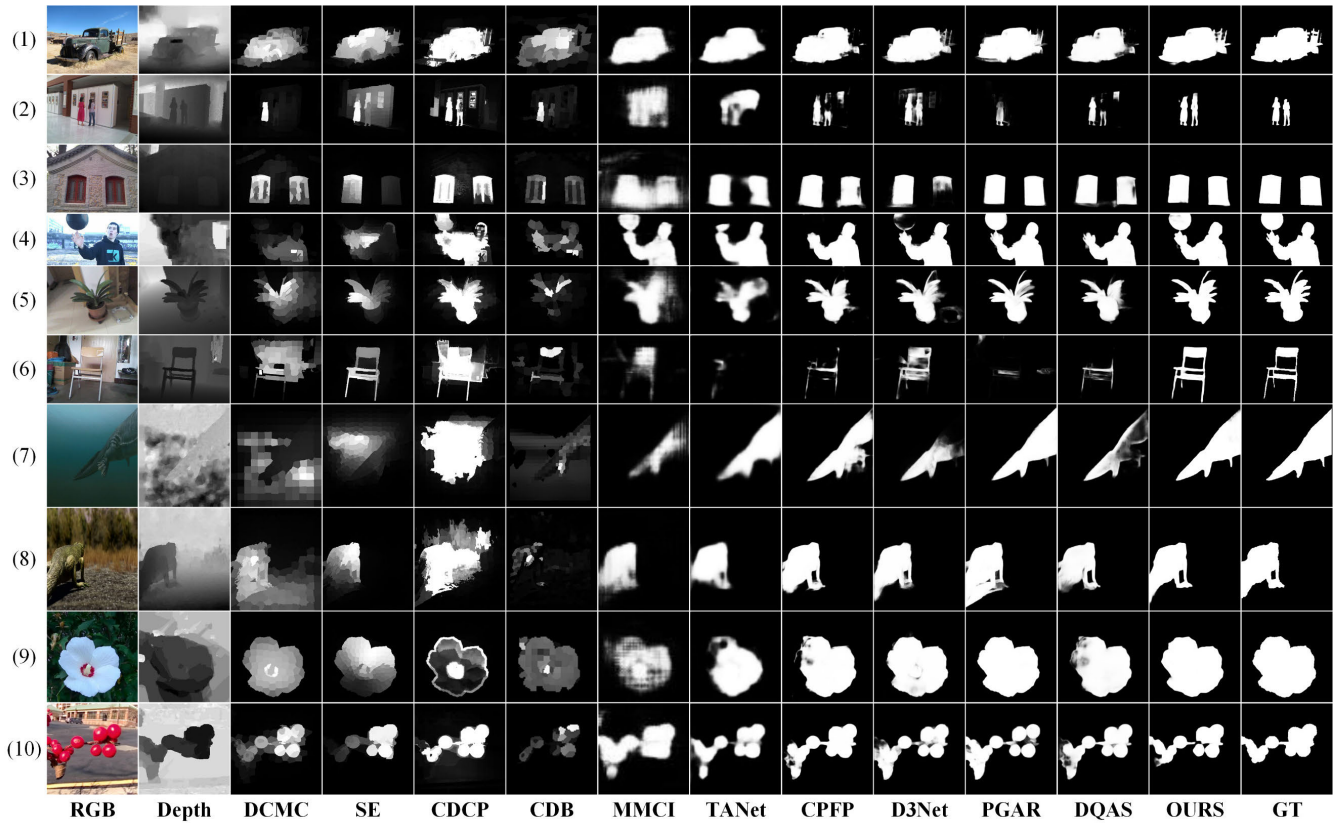


FIGURE 8. Visual comparisons with four latest non-deep learning methods (i.e., DCMC [22], SE [12], CDCP [14] and CDB [16]) and six latest CNN-based methods (i.e., MMCI [26], TANet [31], CPFP [28], D^3Net [30], PGAR [38] and DQAS [43]).

the 3^{rd} and 7^{th} rows, which makes it difficult to distinguish the salient objects by relying on the depth map solely. Moreover, only ours and PGAR algorithms show better performance. Meanwhile, we further find that PGAR responds to objects similar with the background very weakly (i.e., the 6^{th} row), which indicates that the algorithm focuses more on RGB modality and not enough on depth information. Finally, when the object in the depth map is interfered by the background object near ground or close distance (e.g., the 1^{st} and 2^{nd} rows), the results of all comparison methods are not ideal and will be false and missed detections. In sharp contrast, our saliency maps obviously possess more perfect and more detailed salient object.

We further analyze the experimental results in Fig. 8 from the data fusion level. According to the quality of the input data, there are the following four situations. (1) The depth map is very poor, it does not reflect the structural information of the salient object at all, and only be distinguished by color image (e.g., the 3^{rd} and 7^{th} rows). (2) Although the depth map is accompanied by distractors, the partial contour structure of the object is very clear compared to the RGB image, and the color information is relatively obvious on the whole (e.g., the 1^{st} and 2^{nd} rows). (3) The quality of the depth map is relatively good, but the object is difficult to distinguish in terms of color appearance (e.g., the 6^{th} and 8^{th} rows). (4) The object in the RGB and depth maps are all obvious (e.g., the 5^{th}, 9^{th}, 10^{th} rows). For the above four cases, our prediction results are pretty ideal. These show that our model is not disturbed by the extremely poor depth map, and it can also extract the structure of the object from the depth map with certain interference which is difficult to extract from the RGB image. Moreover, our model can make full use of the structure information from the high-quality depth map, no matter whether the color information of the object is prominent or not. All these phenomena indicate that our model is very flexible and effective.

In conclusion, our algorithm is more robust and adaptable in various complex scenarios. It is more inclined to integrate multi-modal cues adaptively and selectively, rather than simply biasing to a certain data branch.

4) OTHER COMPARISONS

In this part, we further analyze the performance of the proposed model in terms of compatibility and model size.

a: COMPATIBILITY

Most of the deep learning-based RGB-D SOD models generally adopt ResNet-50 [45] or VGG-16 [46] as backbone architecture. Therefore, to verify the compatibility of our model, we provide performance comparisons employing different backbones in Table 4. Meanwhile, we utilize diversiform color marks to better reflect the advantages of using ResNet-50 or VGG-16 as backbone compared with other

TABLE 4. Performance comparison using different backbones. Red indicates the best. In PGFNet (VGG-16), bold and green indicate the best and second performances compared with the comparison methods in Table 2. ↑ denotes larger is better, and ↓ denotes smaller is better.

Models	NJU2K [17]				NLPR [9]				STERE [5]				DES [8]				SSD [13]				SIP [30]				LFSD [52]			
	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	$M\downarrow$
PGFNet (VGG-16)	**0.9121**	**0.9465**	**0.9121**	**0.0372**	**0.9195**	**0.9580**	**0.9095**	**0.0245**	0.8976	**0.9395**	**0.8950**	**0.0427**	**0.9359**	**0.9761**	**0.9353**	**0.0179**	0.8553	0.9102	0.8310	0.0521	**0.8802**	**0.9223**	**0.8904**	**0.0504**	**0.8802**	**0.9223**	**0.8904**	**0.0504**
PGFNet (ResNet-50)	**0.9153**	**0.9466**	**0.9159**	**0.0341**	**0.9272**	**0.9647**	**0.9185**	**0.0218**	**0.9019**	**0.9436**	**0.9010**	**0.0396**	0.9335	0.9720	0.9284	**0.0162**	**0.8682**	**0.9149**	**0.8532**	**0.0485**	**0.8823**	**0.9262**	0.8896	**0.0480**	0.8346	0.8735	0.8405	0.0877

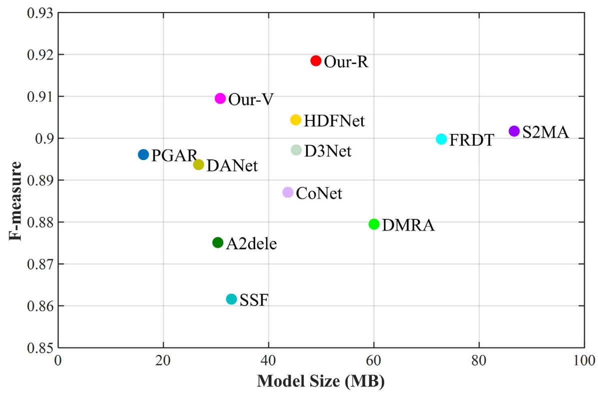


FIGURE 9. Comprehensive comparisons between model size and F-measure performance on the NLPR dataset. Our-R and Our-V denote our models based on ResNet-50 and VGG-16, respectively.

20 state-of-the-art models in Table 1. Obviously, we can discovery that our model is the best overall regardless of the backbone type used, which shows that our framework has strong compatibility.

b: MODEL SIZE

In Fig. 9, we compare the model size of different methods and corresponding maximum F-measure on the NLPR dataset. Compared with other models, our model based on VGG-16 achieves better accuracy with a smaller model size (30.80M). Our model based on ResNet-50 with an acceptable size (48.98M) significantly improves the detection accuracy. The results illustrate that the designed model can realize satisfactory saliency detection performance with a relatively small number of parameters, achieve a certain balance between lightweight and accuracy.

IV. CONCLUSION

In this paper, a novel progressive guided fusion network (PGFNet) is proposed for RGB-D salient object objection. PGFNet is based on a symmetrical two-stream encoder-decoder architecture, which is equipped with three highly efficient submodules with a clear division of labor. Specifically, the multi-modal and multi-scale attention fusion model (MMAFM) to obtain optimal fusion features via learning the internal relationships at different scales and modalities. The multi-modal feature refinement mechanism (MFRM) is applied to enhance the unfused shallow-layer features, and it is achieved through the guidance of the high-level fusion feature. Moreover, the residual prediction module (RPM) to further restrain the background noise according to saliency value. Extensive experimental results on eight datasets prove that our method outperforms most of the state-of-the-art algorithms.

In addition to RGB-D saliency detection, we will consider extending the proposed model to other multi-source data detection or special scenario applications, such as RGB-T salient object detection and underwater computer vision tasks. The RGB-T salient detection [66], [67] integrates RGB and thermal infrared data. Due to the insensitivity of thermal infrared image to light, it can provide supplementary information when the salient object suffers from varying light, glare, or shadows. Furthermore, in relatively recent years, saliency detection has been widely used in the field of automated underwater exploration [68]. In view of this, we consider using dark channel prior [69] instead of depth information as a powerful auxiliary to improve the performance of underwater detection. Theoretically, these schemes are feasible and worthy of further research in the future.

REFERENCES

[1] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[2] R. Achanta and S. Susstrunk, "Saliency detection for content-aware image resizing," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 1005–1008.

[3] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.

[4] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 10, 2020, doi: 10.1109/TPAMI.2020.3023152.

[5] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 454–461.

[6] A. Ciptadi, T. Hermans, and J. Rehg, "An in depth view of saliency," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 112.1–112.11.

[7] K. Desingh, K. M. Krishna, D. Rajan, and C. V. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 98.1–98.11.

[8] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, pp. 23–27.

[9] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 92–109.

[10] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 454–458.

[11] H. Sheng, X. Liu, and S. Zhang, "Saliency analysis based on depth contrast increased," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 1347–1351.

[12] J. Quo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, Jul. 2016, pp. 1–6.

[13] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 3008–3014.

[14] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 1509–1515.

[15] C. Zhu, G. Li, X. Guo, W. Wang, and R. Wang, "A multilayer backpropagation saliency detection algorithm based on depth mining," in *Computer Analysis Images and Patterns*, M. Felsberg, A. Heyden, and N. Krüger, Eds. Cham, Switzerland: Springer, 2017, pp. 14–23.

[16] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, "Stereoscopic saliency model using contrast and depth-guided-background prior," *Neurocomputing*, vol. 275, pp. 2227–2238, Jan. 2018.

[17] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 1115–1119.

[18] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 25–32.

[19] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2343–2350.

[20] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, "Improving RGBD saliency detection using progressive region classification and saliency fusion," *IEEE Access*, vol. 4, pp. 8987–8994, 2016.

[21] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, May 2017.

[22] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Apr. 2016.

[23] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and A. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[24] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.

[25] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[26] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

[27] H. Chen, Y.-F. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Madrid, Spain, Oct. 2018, pp. 6821–6826.

[28] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3927–3936.

[29] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7254–7263.

[30] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[31] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[32] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for RGB-D image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, Oct. 2019.

[33] H. Chen, Y. Deng, Y. Li, T. Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, 2020.

[34] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2020, pp. 665–681.

[35] M. Zhang, Y. Zhang, Y. Piao, B. Hu, and H. Lu, "Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4107–4115.

[36] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2020, pp. 646–662.

[37] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[38] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2020, pp. 520–538.

[39] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[40] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[41] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Comput. Vis. Media*, vol. 7, pp. 37–69, Jan. 2021.

[42] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusion scheme for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 458–471, 2021.

[43] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Depth quality-aware selective saliency fusion for RGB-D image salient object detection," *Neurocomputing*, vol. 432, pp. 44–56, Apr. 2021.

[44] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4296–4307, 2020.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[47] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[49] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[50] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 7012–7024, 2021.

[51] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12321–12328.

[52] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[53] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4558–4567.

[54] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[55] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

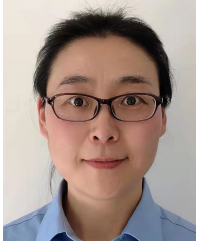
[56] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

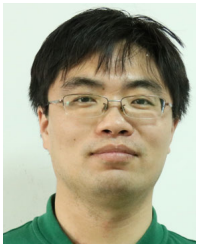
[58] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 275–292.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[60] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3051–3060.
[61] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3472–3481.
[62] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.
[63] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 235–252.
[64] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2020, pp. 52–69.
[65] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9060–9069.
[66] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Proc. Chin. Conf. Image Graph. Technol.* Singapore: Springer, 2018, pp. 359–369.
[67] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, Jan. 2020.
[68] M. Reggiannini and D. Moroni, "The use of saliency in underwater computer vision: A review," *Remote Sens.*, vol. 13, no. 1, p. 22, Dec. 2020.
[69] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

JIAJIA WU received the B.S. degree from Northeastern University at Qinhuangdao, in 2017. She is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. Her current research interests include RGB-D salient object detection and deep learning.

GUANGLIANG HAN received the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2000 and 2003, respectively. He is currently a Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His current research interests include computer vision, image processing, and object tracking.



HAINING WANG is currently working with the School of Police Administration, People's Public Security University of China. From 2010 to 2011, she was a Visiting Scholar with Ohio University. Her research interests include scientific decision making, public security informationization, and information theory.



HANG YANG received the B.S. and Ph.D. degrees from Jilin University, in 2007 and 2012, respectively. He is currently an Associate Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include image restoration and object tracking.

QINGQING LI received the B.E. degree from Hainan University, China, in 2017. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include image registration, image fusion, and deep learning.

DONGXU LIU received the B.E. degree from the Nanjing University of Information Science and Technology, China, in 2018. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include hyperspectral image classification and deep learning.

FANGJIAN YE received the M.A. degree in 2012. He has been working as a forensic scientist for 16 years. He is currently an Associate Research Fellow with the Institute of Forensic Science, Ministry of Public Security, China. His research interests include crime scene investigation, footprint inspection and identification, and tool marks examinations.

PEIXUN LIU received the Ph.D. degree from Jilin University, in 2015. He is currently an Associate Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include image processing, object detection, and robot automation.